Educational Big Data Mining: Comparison of Multiple Machine Learning Algorithms in Predictive Modelling of Student Academic Performance

Educational Big Data Mining

Ting Tin Tin[1], Lee Shi Hock[2], Omolayo M. Ikumapayi[3]
Faculty of Data Science and Information Technology, INTI International University, Nilai, Negeri Sembilan, Malaysia[1]
Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia[2]
Department of Mechanical and Industrial Engineering, University of Johannesburg, Johannesburg, South Africa[3]

Abstract—**Utilisation of Educational Data Mining (EDM) can be useful in predicting academic performance of students to mitigate student attrition rate, allocation of resources, and aid in decision-making processes for higher education institution. This article uses a large dataset from the Programme for International Student Assessment (PISA) consisting of 612,004 participants from 79 countries, supported by the machine learning approach to predict student academic performance. Unlike most of the literature that is confined to one geographical location or with limited datasets and factors, this article studies other factors that contribute to academic success and uses student data from various backgrounds. The accuracy of the proposed model to predict student performance achieved 74%. It is discovered that Gradient Boosted Trees surpass the other classification models that were considered (Logistic Regression, Naïve Bayes, Deep Learning, Random Forest, Fast Large Margin, Generalised Linear Model, Decision Tree and Support Vector Machine). Reading skills and habits are of the highest importance in predicting the academic performance of students.**

Keywords*—*Academic performance; CGPA; education data mining; machine learning; predictive modelling; R&D investment

I. Introduction

Improving student academic performance to create a sustainable and quality education ecosystem has been the primary goal of Higher Education Institutions (HEIs). However, the identification of factors that influence academic performance is still lacking in completeness. This has resulted in an increase in the interest in research to extract knowledge from the accumulated data to identify the significant factors that influence academic performance [1,43]. This can be achieved through data mining methods and techniques commonly known as Educational Data Mining (EDM). The goal of EDM is to provide methods for examining the distinctive types of data generated by educational environments to aid educational activities and predict academic performance [2,8,37]. Due to EDM's functionality, it has become an emerging field of research to extract critical knowledge from existing students' data to help in decision-making processes. The motivation behind this research is to employ data analytics to identify low-performing students to provide timely intervention through academic guidance to overcome their learning obstacles, thus improving their learning outcomes.

Although most HEIs are aware of the potential of EDM, they have not been able to analyse this data and turn it into useful information. Even in HEIs that apply EDM, its application of it is suboptimal as traditional methods backed by statistics are used [27]. Predicting student academic performance is crucial for institutional leaders, students and their families, and policy makers [18]. This is because student academic performance acts as a measure of the institution's success [4]. Increasing performance is in tandem with the Ministry of Higher Education (MoHE) effort to transform Malaysia into a regional hub for higher education. However, since the outbreak of Covid-19, there has been a sharp increase in student attrition rates [44]. Berens et al. emphasised that student attrition is a misuse of the allocated public and private resources [11]. Therefore, it is necessary to implement this predictive model to understand the factors that contribute to academic failure and provide timely intervention [43]. Thus, the success rate of the students would be significantly improved.

The use of data mining techniques to predict academic performance of students has been adopted in many studies [1,25,27,34]. However, most studies are restricted to a particular geographic area due to limited data sources or restricted to a few factors that influence academic performance. In previous studies, the deficiencies of limited datasets have been acknowledged to be insufficient in accurately predict academic performance [31]. Furthermore, geographically constrained sets would cause concerns about external validity in the usage of external regions, which impacts the practicality of the predictive model used. In response to these issues, this study opts for large datasets (600k participants consisting of international groups of students' academic performance from 79 countries) inclusive of a multitude of factors (demographic, family background, learning environment at home and school, reading skills and habits, career goals and mindset, mental health, lifestyle, and IT knowledge) to predict students' academic performance obtained from Programme for International Student Assessment (PISA).

The objective of this article is to fill the gaps in the existing literature in predicting the prospective academic performance of a student through classification algorithms. Input factors will also be screened to decide which factors have a high influence on academic performance. The algorithm with the highest accuracy will be chosen to integrate into web services. The following research questions are proposed:

1) Which data mining techniques have the highest accuracy in predicting students' academic performance?

2) What factors are significant in predicting students' academic performance?

II. Literature Review

With the abundance of student data available through institutional education systems, there has been a tremendous increase in interest in the use of data mining techniques in educational settings [27]. Educational data mining (EDM) has become a vital field for extracting knowledge from existing student data to aid teaching and learning activities [37]. Through this, student performance could be successfully predicted, reducing student attrition rates and helping educational institutions make managerial decisions [43]. The application of EDM is not limited to these two aspects; it can also be applied to predict student dropout [16,32,36] and develop recommender systems [15,38]. Among all EDM applications, its most prominent application is predicting student performance [27]. However, this is a challenging task due to the intricate nature of factors that can affect student performance [27], the large amount of data in educational databases [41], the diverse backgrounds of students, and the inadequate comprehension of the skills required for success [45].

Taking into account the scenarios, research on EDM to predict students' academic performance focusses only on certain attributes to characterise students and their environments. For example, a study by Yakubu and Abubakar [43] predicts academic performance with demographic information and high school graduation exam scores as input variables. Meanwhile, a study by Roy et al. [34] only includes CGPA as an attribute. Fernandes et al. [19] use three attributes, namely demographic information, past grades, and environment. It should be noted that, based on a survey conducted by Abu et al. [1], of 36 of 420 articles that were critically reviewed and conducted in the period from 2009 to 2018, most of the scope of the research focusses on students' learning activities (25%), past academic grades (26%) and demographics (23%). Abu et al. [1] also argued that more studies are required to identify the factors that influence student performance. Taking into account this situation, it is critical to analyse other factors that present a correlation with academic performance. After extensive research, factors to be considered are demographic information, family history, home learning environment, reading skills and habits, career goals and mindset, and mental health of students.

A. Demographic

Fernandes et al. [19] examined how student demographics, such as their neighbourhood, type of school, city condition and age, are highly associated with their academic performance. However, when considering the students' grades, the significance of demographic variables reduces. Silva et al. [40] analysed demographics in terms of gender, ethnicity, medium of instruction, and differences in school category in science performance. They reported that female students have higher academic performance than their male counterparts. Their findings also showed that schools with better infrastructure outperformed those with poorly equipped infrastructure. Well-equipped schools tend to receive better allocation and attract more academically inclined students. There is also a clear disparity in science performance based on the medium of instruction, which is correlated with ethnicity.

B. Home Learning Environment

Since the start of the pandemic, teachers and students have adopted online learning [10]. However, the factor of technology poses a concern, particularly for marginalised or remote students, as online learning discriminates against them. Due to poor internet reception and lack of electricity, students from these backgrounds tend to have unfavourable learning environments at home, preventing them from participating in the online learning process. The authors found that many students lack a conducive study environment at home, affecting their learning process. Poverty exacerbates this situation, as students lack the digital resources needed for learning activities, resulting in poor attendance in online classes. These circumstances are more likely to cause students to drop out of school [22]. García-González and Skrita [20] further support Kapasia's findings [22], revealing that owning a personal computer and having a higher socioeconomic status positively impact academic results.

C. Family Background

Family background refers to the conditions and circumstances that impact the physical, intellectual and emotional development of a child [29]. This sentiment is echoed by Li and Qiu [46], who suggest that families play a critical role in shaping children's learning behaviours, which, in turn, impacts academic achievement. This notion is supported by research, which indicates that children's educational attainment is significantly influenced by the level of educational investment made in them. This is particularly challenging for families from low socioeconomic backgrounds, who may not be able to invest adequately in their children's education, which ultimately affects their academic achievement. In contrast, families with rich cultural capital can provide the necessary resources, environment and support to cultivate their children's motivation and interest in academic pursuits, allowing them to perform better in their academics [46]. In addition, families with higher social and economic status tend to have a positive correlation with the quality of education their children receive. Parents of these backgrounds are more likely to have higher educational backgrounds and are better equipped to secure quality educational opportunities for their children. This statement is supported by García-González and Skrita [20], who found a positive correlation between higher socioeconomic status and educational level in the family and higher academic performance. Therefore, students with better resources, a conducive environment, and attentive parents tend to have a positive association with academic success [21,46].

D. Reading Habits

Reading is the ability to extract meaning from words written in textual or digital form to seek knowledge, information, or entertainment [9]. Kumara and Kumar [23] stipulate that reading is not solely the process of deriving information through text, but a multifaceted journey that combines the intellect of the greatest thinkers of all time. Reading provides readers with a gateway to form their own understanding of the subject and promotes novel ideas. Therefore, reading is an activity that involves evaluation, judgment, foresight, and critical thinking. Reading helps to develop logical thinking and the creation of new ideas, as supported by Le et al. [24], who suggest that avid readers have a better development of critical thinking and adaptability skills. Therefore, students with undeveloped reading habits tend to perform poorly academically and have a greater propensity to engage in delinquent acts [35]. The authors further argue that students with poor reading habits are left behind during class activities, and this cycle continues throughout their academic lifetime. As a result, reading ability, which is related to reading habits, is a determinant of academic performance [24].

E. Career Goal and Mindset

In short, career goals, mindset, or career aspirations are regarded as the dreams, desires, and ambitions that people aspire to pursue in a particular career field by joining relevant courses. This enables individuals to pinpoint and create goals through the contextualisation of present and past perspectives. Therefore, career aspirations help people to be inspired and actively pursue their aspirations [26]. Career goals are vital to an individual's overall development because they act as a guide for academic and career achievements while navigating adulthood. This is mainly because career development, formed largely through career aspirations, has a significant impact on the development of one's intellect, emotions, and social skills [33].

Le et al. [24] noted that students with high aspirations to work in the fields of Science, Technology, Engineering, and Mathematics (STEM) tend to have better academic performance in STEM subjects. These students actively pursue co-curricular activities related to STEM. Thus, a positive association with better performance in STEM subjects could be observed in students who intend to continue in STEM-related fields. The views of Le et al. [24] are further supported by Margaret [26], who described that career aspirations shape the study behaviour of an individual. As a result, it would instil a growth mindset, which translates to an improvement in academic performance.

F. Mental Health

It is quite concerning that one in three students will not complete their tertiary education successfully (Organisation for Economic Cooperation and Development (OECD) [30]. This problem persists due to the lack of awareness of mental health among students. Research stipulates that students who experience difficulties in adjusting to university life are more prone to academic failure. The difficulty in adapting to the transitional phase in college is related to the transition from late adolescence to emerging adulthood, which is a critical period due to the shift in autonomy by having to be self-reliant and independent. This situation can lead to a higher prevalence of mental health problems, leading to low levels of satisfaction with life [5,12,17].

Students who suffer from mental health problems are established to experience an existential crisis, where they have no particular purpose in life [39]. This issue is further exacerbated by the insurmountable stress of academic underperformance. Cant [14] stated that study stress is correlated with the onset of mental health problems such as anxiety, low self-esteem, and depression. His findings are supported by Agnafors, Barmark, and Sydsjö [6], where low academic performance among 16-year-old students is associated with depression. Therefore, it is crucial to provide early intervention to mitigate mental health problems in early childhood and adolescence, as they are interrelated with academic performance.

A review of the literature shows that there is potential for growth in several domains. First, most of the literature conducted is strictly focused on a specific geographical location of educational institutions, combined with a limited number of factors [31]. Therefore, there is room for improvement in the use of datasets. This is because confining the research to a specific geographic area would cause generalisability concerns in the predictive model. The limited use of factors would also affect the precision of the predicted academic performance [42]. Second, most of the literature considers a limited number of data sets [27]. This is noted by Oyedeji et al. [31], where limited datasets would limit data mining tools to make more accurate predictions about academic performance. As a result, this would undermine the information extracted and thus impact the practicality of the predictive model used.

G. Machine Learning Classification Techniques

Nine popular data mining classification algorithms are used to predict the academic performance of students. Logistic Regression (LR), Naïve Bayes (NB), Generalized Linear Model, Fast Large Margin, Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees, and Support Vector Machine (SVM) (Table I) [47, 48]. According to Abu et al. [1], the NB and DT classifiers, along with Artificial Neural Networks, which fall under DL, are the most frequently used data mining algorithms. This statement is also supported by Miguéis et al. [27], where DT and NB are popular in the context of EDM due to their performance and the efficiency of the training effort. Therefore, testing other data mining algorithms can contribute significantly to EDM research. In addition, the different dependent variables (DV) used in predictive modeling also affect the performance of different algorithms. Therefore, different algorithms are required in the predictive modelling testing for different contexts (Table I).

III. Methodology

The proposed framework, as illustrated in Fig. 2 consists of six stages: 1) Data Collection, 2) Initial Preparation, 3) Statistical Analysis, 4) Data Preprocessing, 5) Data Mining Implementation, and 6) Evaluation. The data set was acquired from the Programme for International School Assessment (PISA) and consists of a student population between the ages of 15 years and 3 months and 16 years and 2 months from 79 participating countries. Students are currently enrolled in an educational institution in grade 7 or higher, from at least 150 schools per country. Exclusions were applied to the data set at both the school and the student levels. The former excluded geographically inaccessible schools and special needs schools, while the latter excluded students with intellectual or functional disabilities and students who demonstrated limited language proficiency in the PISA testing environment. Data pertain to the eligible student population and include a total of 612,004 respondents from both computer-based and paper-based tests. It comprises academic information on students' performance in reading, mathematics, and sciences, as well as their global competence, well-being, demographics, ICT familiarity, and financial literacy.

A. Initial Preparations

The original data obtained are in raw format, which is not suitable for analysis and modelling because of its unstructured nature. This is because the data may be inconsistent, incoherent, incorrect, incomplete, contain duplications, or include noise such as errors and outliers. Therefore, the raw data must undergo initial preparation, which is divided into three stages: 1) data selection, 2) data cleaning, and 3) data derivation. Data transformation is the most crucial process to ensure the conversion of raw data from an unstructured format into a structured, comprehensible, and precise format.

TABLE. I. RECENT TWO-YEAR STUDIES ON PISA DATASET WITH MACHINE LEARNING ALGORITHMS IN ANALYTICS

DV(s)	IV(s)	ML algorithm used	Best algorithm	Performance	Resource
Life's satisfaction	meaning in life, student competition, teacher support, exposure to bullying, ICT resources at home and at school	RF, KNN	RF	RMSE=.451	[49]
Science	ICT use, demographics, parents, class discipline, well-being, learning time, socio-economic, teacher	XGBoost, LR, SVR, RF	XGBoost	RMSE=79.96	[50]
Mathematics	Demographic, socio-economic	RF, LR, SVM,	SVM	Accuracy=.797	[51]
Science	literacy, parents' educational status, the disciplinary environment in the classroom, learning time	SVM	SVM	Accuracy=0.78	[52]
Academic performance	Global competence, gender, public/private school	boosted-regression-tree	boosted-regression-tree	R^2=0.75 RMSE=47.74	[53]
Science	Science context, knowledge, competencies, background, home, school, learning experiences	SVM, LR, MLP, DT, RF	RF	Accuracy=0.74	[54]
Reading	Student, teacher, school	MLM	MLM	-	[55]
Reading	personal characteristics, proximal processes, contextual factors	RF, HLM	RF, HLM	RF: $R^2 = 0.66$; RMSE = 47.19 HLM: $R^2 = 0.64$; RMSE = 47.50	[56]
Digital reading	Individual, home, school	SVM	SVM	Accuracy=87.51	[57]
Reading	Metacognitive strategies, reading interests	GBDT	GBDT	RMSE > 65.7	[58]
Science, Mathematics	Well-being	Boosted tree, neural boosted, XGBoost, Bootstrap forest	Neural boosted	Mean RASE=78.12	[59]
Well being	Bioelogical – individual, proximal process, context	GBDT, AdaBoost, ET, RF, LightGBM	LightGBM	MAE=0.342-1.557	[60]
Reading	Teacher, school, parents	DT, NB, KNN, RF	RF	Hamming score=.8427	[61]

Note: DV-dependent variable; IV-independent variable; RF-random forest; KNN-k-nearest neighbours; LR-logistic regression; SVR-support vector regression; XGBoost-extreme gradient boosting; GBoost-gradient boosting; SVM- support vector machine; MLP- multilayer perceptron; DT-decision tree; MLM-multilevel model; HLM- hierarchical linear modelling; GBDT-gradient boosted decision tree; AdaBoost- adaptive boosting; ET-extratrees; LightGBM- light gradient boost machine; NB-Naïve Bayes

Data Selection: The size of the acquired data is important as it consists of several attributes that could negatively affect computational complexity. However, using all the acquired data in the analysis phase would produce suboptimal predictions in the event of data dependency or redundancy. To avoid this, attributes that significantly impact the prediction results need to be determined and included in the analysis phase by understanding the EDM goals and the data itself. Doing so would prevent overfitting, as the predictive model would not be fed with excessive data (i.e., data with a high number of features). Data selection or dimensionality reduction is a technique that consists of vertical selection of attributes and horizontal selection of instances to reduce the number of features in the data set, thus avoiding the Curse of Dimensionality and providing the predictive model with an optimal data set [4].

Data Cleaning: The original data set often contains missing values, inconsistencies, and noises. Missing values occur when a value is not present for a variable, and outliers occur when a value deviates significantly from other values. Therefore, these occurrences need to be cleaned without compromising the efficacy of the prediction model. If left untreated, missing values could compromise the quality of some classifiers, namely Support Vector Machines (SVM), Naïve Bayes (NB), Neural Networks (NN), and Logistic Regression [4]. However, Random Forests and Decision Trees can handle missing data [3]. In this study, two strategies are implemented to resolve missing values as shown in Fig. 1. The first strategy is by list deletion, where either the record (row) or the variable (column) is deleted if the missing value percentage exceeds 50%. The next strategy is by imputation, where the missing value(s) is derived from the remaining data (mean, median, constant value for numerical value or random value from the distribution of missing values) [28].

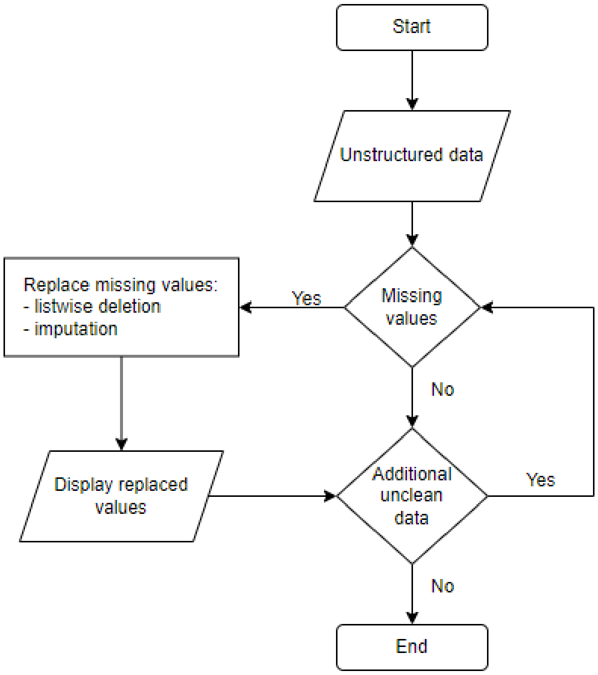


Fig. 1. Data cleaning flow chart for the PISA dataset.

To determine whether the selected variable has missing values, a statistical analysis is performed using descriptive statistics. The variable ST004D01T will be taken as an example. From the descriptive statistics result, there are a total of 612,002 valid data and 2 missing values. The value 1 indicates Female, while the value 2 indicates Male. Since missing values are present and the percentage of missing values is less than 50%, imputation will instead be used as a list deletion. Table I illustrates the result of the statistical analysis before data cleaning is performed. In this case, the missing value is replaced using the mode value 2. The missing value is now replaced by the male category. A similar approach is applied to the other variables until all the data have been fully cleaned.

Data Derivation: Data derivation is a process in which new variables are derived from existing variables by constructing or combining them. The combination of different variables is done when they possess similarities. Applying a data derivation based on domain knowledge would improve the data mining system. For example, the variable "age" is derived from the "date of birth" (DOB) attribute. If taken as it is, the DOB does not explicitly state the student's age. This analogy could be applied in the context of parents' education levels. Two pieces of data will be analysed and combined to determine the level of education level. Subsequently, this would provide more in-depth information about the student's family background, as opposed to taking the information from a standalone viewpoint.

Data Pre-Processing: Data preprocessing is the final step before data analysis and modelling could be performed. This step consists of 1) Data Transformation and 2) Feature Selection.

Feature selection: After successful transformation of the dataset, it is now ready to undergo modelling by selecting important variables and inserting them into the modelling algorithm. This is a crucial step that must be done before data mining can proceed. Feature selection enables reduction of computational complexity, computation time, and enhancement of prediction performance, as well as better comprehension of the data. This can be achieved through filter or wrapper methods. The filter method is a type of preprocessing stage that aims to rank and identify features that would significantly affect the prediction result before applying them to the prediction model. Wrapper methods, on the other hand, involve wrapping the predictor onto the algorithm to identify the features that would provide the best prediction result.

B. *Data Mining*

There are two types of data mining models that are applied in EDM applications: predictive and descriptive models. For this project, a predictive model will be used. Predictive models use supervised learning methods to estimate the expected values of dependent variables based on the characteristics of the respective independent variables. Predictive models can be classified into classification and regression methods [13]. It should be noted that classification is the most widely used technique followed by regression. Common examples of classification techniques are Decision Trees, NN, and Bayesian Networks, while regression includes linear and logistic regression, respectively. When deciding on the data mining model, the algorithms to build the predictive model are considered based on their performance and accuracy. The algorithm with the best performance is chosen before making any configurations in further stages. This includes using the trial-and-error method by fine-tuning its parameters to increase its efficiency. Subsequently, the parameters that provide the most effective performance are chosen before application [4]. Various open source tools are available for data mining, helping researchers analyse data sets using built-in algorithms. These tools are widely used for visualisation, predictive analysis, and modelling. Therefore, RapidMiner and IBM SPSS will be used due to its built-in functionality for preprocessing, association rules, classification, regression, and visualisation, as well as its accessibility.

C. *Evaluation*

The prediction model's performance will be assessed using the confusion matrix which consists of two classes: the predicted and the actual class. Different performance measures (accuracy, precision, and recall) could be determined to evaluate the performance of each prediction model with the respective algorithms used. Fig. 2 summarises the three phases of this research activities to produce predictive modelling of academic performance from machine learning.

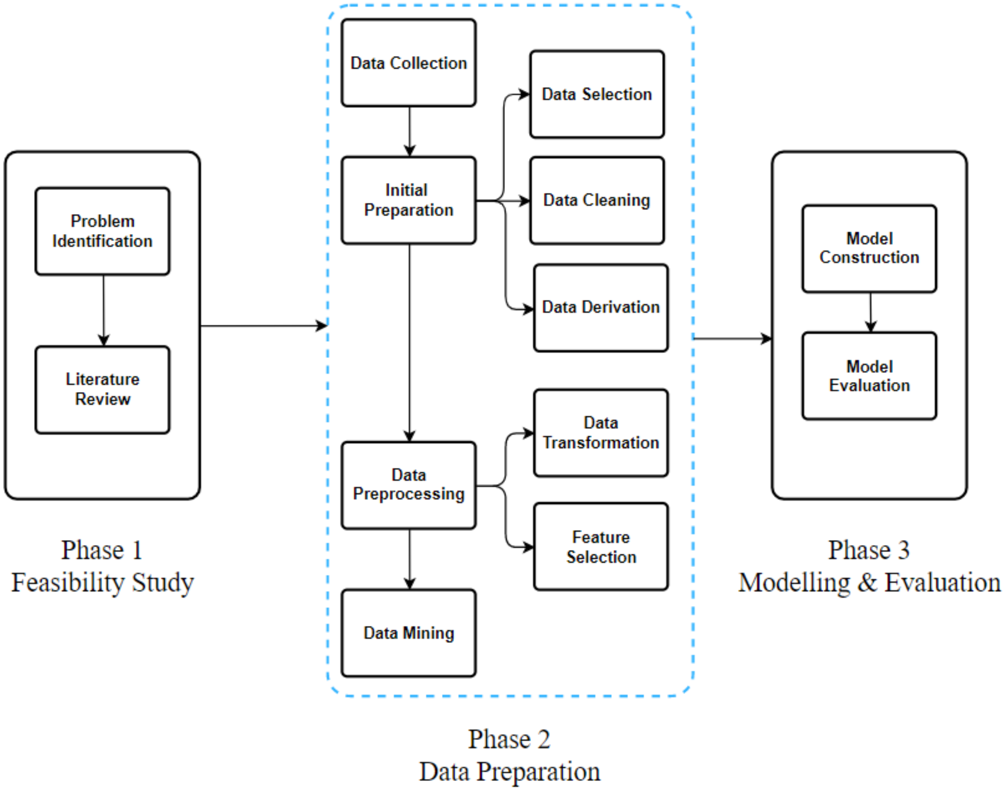


Fig. 2. Proposed activities to predict student academic performance using a classification model.

IV. Result and Discussion

A. CGPA Discretization

Four levels of academic performance are established through a discretization algorithm. Academic performance (AP) is categorised into four groups - A, B, C, and D, according to plausible values (PV) from the PISA dataset. Four categories of PV, Mathematics, Reading, Science, and Global Competency, are used to calculate the AP. Each category of PV consists of 10 values. Data derivation is performed by taking the first PV for each category and then averaging them to compute the AP, which ranges from 157 to 809. The difference between the highest and lowest values is computed and then divided by four. The resulting value is added to the lowest value and is performed for each level. The resulting AP is shown in Table II, which highlights the percentage and range for each group.

TABLE. II. Percentage and Range of Academic Performance

Academic Performance	Range	Percentage of students
A	157 – 319	3.34
B	320 – 482	55.61
C	483 – 645	39.86
D	646 - 809	1.19

B. Attribute Selection

Before modelling the algorithm, attributes were selected based on their predictor importance regarding academic performance (AP), from the identified factors: demographics, learning environment at home, family background, reading skills and habits, career goals and mindset, and mental health of students. This was done by ensuring that the questions were related to the identified factors. The list of questions for each selected attribute is included in Appendix A1. Multiple regression analysis was used to compute the R value and significance of the variables chosen to predict AP. Linear regression was also used to derive the importance of the predictor for each variable considered in the model. This enabled us to gain an understanding of the importance of each characteristic in predicting a student's academic performance. The sum obtained was normalised to bring the values into the range [0,1].

Only one demographic variable was identified, Gender. As the PISA dataset only includes students who are 15 years old and from a high school background, age could not be considered as one of the variables. The data set available for school category differences is not in line with the quantity of data available for student categories by 590k. Furthermore, the predictor variables used do not consider differences in school infrastructure and school categories. Therefore, school-related variables were not examined for their relationship with academic performance. R-squared (R^2) measures how well IV explains the variation in DV. When analysing the value of the results of Table III, the R^2 is 0.04, which is < 0.3. Therefore, the variable has a moderately weak effect size on AP. However, the significance value in Table IV is 0.00, which implies that it does affect the prediction of a student's AP. Since there is only one variable, the importance of the predictor is 1.00 as shown in Table V. Besides that, all the predictors show significant impacts on AP (Table IV) with respect to the AP.

TABLE. III. MODEL SUMMARY

Variable Category	R	R^2	Adjusted R^2	Std. Error
Demographic	0.059	0.004	0.04	82.41
Home learning envi	0.354	0.125	0.125	77.20
Family Background	0.336	0.113	0.113	77.75
Reading skills and habits	0.587	0.345	0.345	66.82
Career goals and mindset	0.336	0.100	0.100	78.32
Mental health	0.206	0.042	0.042	80.79

TABLE. IV. ANOVA TEST

	SS	df	MS	F	Sig.
Demographic					
Regression	14710294	1	14710292.98	2166.03	0.00
Residual	4156322639	612002	6791.355		
Total	4171032933	612003			
Home Learning Environment					
Regression	5231815912	5	10463618.4	17554.78	0.00
Residual	3647851341	611998	5960.561		
Total	4171032933	612003			
Family Background					
Regression	471500233	17	27735307.81	4588.044	0.00
Residual	3699532700	611986	6045.13		
Total	4171032933	612003			
Reading skills and habits					
Regression	1438720511	40	35968012.78	8055.85	0.00
Residual	2732312422	611963	4464.83		
Total	4171032933	612003			
Career goals and mindset					
Regression	416708462	21	19843260.09	3234.60	0.00
Residual	3754324471	611982	6134.69		
Total	4171032933	612003			
Mental Health					
Regression	176266901	17	10368641.22	1588.44	0.00
Residual	3994766032	611986	6527.55		
Total	4171032933	612003			

Note: SS- sum of squares; df-degrees of freedom; MS-mean sum of squares

TABLE. V. PREDICTOR IMPORTANCE

Variables	Significance	Importance
Demographic		
ST004D01T	0.000	1.00
Home learning environment		
ST011Q04TA	< 0.001	0.53
ST011Q06TA	0.000	0.26
ST011Q01TA	0.000	0.20
ST011Q03TA	< 0.001	0.01
ST011Q02TA	0.000	0.00
Family Background		
MISCED	0.000	0.29
ST123Q02NA	0.000	0.28
FISCED	< 0.001	0.20
PA009Q09NA	< 0.001	0.07
EC155Q01DA	0.000	0.04
PA008Q09NA	< 0.001	0.02
PA008Q05TA	< 0.001	0.02
EC155Q02DA	< 0.001	0.02
ST123Q04NA	< 0.001	0.01
PA008Q03TA	< 0.001	0.01
Reading skills and habits		
ST013Q01TA	0.000	0.37
ST154Q01HA	< 0.001	0.08
ST161Q06HA	< 0.001	0.07
ST152Q05IA	< 0.001	0.05
ST153Q06HA	0.000	0.04
ST153Q09HA	< 0.001	0.03
ST160Q05IA	< 0.001	0.03
ST161Q03HA	< 0.001	0.03
ST011Q08TA	< 0.001	0.03
ST160Q04IA	< 0.001	0.02
ST011Q11TA	< 0.001	0.02
ST011Q07TA	< 0.001	0.02
ST153Q01HA	< 0.001	0.02
ST160Q02IA	< 0.001	0.02
ST011Q12TA	< 0.001	0.02
ST153Q05HA	< 0.001	0.02
ST161Q02HA	< 0.001	0.01
ST167Q04IA	< 0.001	0.01
ST168Q01HA	< 0.001	0.01
ST150Q03IA	< 0.001	0.01
ST160Q01IA	< 0.001	0.01
ST167Q02IA	< 0.001	0.01
ST153Q03HA	< 0.001	0.01
ST167Q03IA	< 0.001	0.01
ST153Q10HA	< 0.001	0.01
ST150Q02IA	< 0.001	0.01
ST167Q05IA	< 0.001	0.01
ST153Q04HA	< 0.001	0.01
Career goals and mindset		
EC152Q01HA	0.000	0.37
EC153Q02HA	0.000	0.13
EC150Q05WA	< 0.001	0.12

EC153Q07HA	< 0.001	0.07
EC150Q06WA	0.000	0.06
EC150Q07WA	< 0.001	0.05
EC153Q01HA	< 0.001	0.05
EC150Q01WA	< 0.001	0.04
EC153Q05HA	< 0.001	0.04
EC153Q06HA	0.000	0.02
EC150Q09WA	< 0.001	0.02
EC153Q04HA	< 0.001	0.01
EC153Q11HA	< 0.001	0.01
EC153Q08HA	< 0.001	0.01
Mental Health		
ST186Q09HA	< 0.001	0.24
ST185Q01HA	< 0.001	0.12
ST186Q01HA	< 0.001	0.09
ST186Q05HA	< 0.001	0.08
ST185Q03HA	0.000	0.08
ST186Q02HA	< 0.001	0.08
ST186Q10HA	< 0.001	0.04
ST186Q08HA	< 0.001	0.04
WB171Q03HA	< 0.001	0.03
ST186Q06HA	0.000	0.03
ST185Q02HA	< 0.001	0.03
ST186Q07HA	0.829	0.03
ST186Q03HA	0.089	0.03
WB171Q02HA	< 0.001	0.02
WB171Q04HA	0.104	0.02
WB171Q01HA	< 0.001	0.02
ST016Q01NA	< 0.001	0.02

A total of five variables were identified for the learning environment at home. The value of R^2 is 0.125, indicating a weak effect size of $R^2 < 0.3$. However, the significance value is 0.00, which implies that the variables are significant in predicting student AP. It is not surprising to observe that having a computer that can be used for schoolwork and Internet connectivity are the two most significant predictors. This suggests that having access to digital devices along with the Internet would significantly boost academic performance by providing the necessary resources to carry out learning activities.

Seventeen variables related to family history were identified. The value of R^2 is 0.113, indicating a weak effect size of $R^2 < 0.3$. However, the significance value is 0.00, which implies that the variables are significant in predicting a student's AP score. Of the 17 variables identified, only 10 are considered important in predicting a student's AP score. Mother's education, parental support toward educational achievements, and father's education claimed the top three spots, respectively. This suggests that parents with a higher educational background prioritise their child's academic success.

Forty variables related to reading skills and habits were identified. It was observed that the R^2 value is 0.345, indicating a weak effect size of $0.3 < R^2 < 0.5$. However, the significance value is 0.00, which implies that the variables are significant in predicting a student's AP score. Only 28 are considered important in predicting a student's AP score. Surprisingly, the number of books at home (ST013Q01TA) has the highest predictor importance in a student's AP score. The number of books at home affects a child's literacy level and serves as an indicator of their socioeconomic status.

Twenty-one variables related to career goals and mindset were identified. The value of R^2 is 0.100, indicating a weak effect size of $R^2 < 0.3$. However, the significance value is 0.00, which implies that the variables are significant in predicting a student's AP score. Interestingly, of the 21 variables, only 14 are considered important in predicting a student's AP score. Students who have a clear idea of their career goals five years from now have the most significant predictor importance, allowing them to contextualise their goals through academic achievement.

Seventeen variables related to career goals and mindset were identified. The value of R^2 was observed to be 0.100, indicating a weak effect size of $R^2 < 0.3$. However, the significance value is 0.00, which implies that the variables are significant in predicting a student's AP score. All variables are considered important in predicting a student's AP score. However, WB171Q04HA and ST186Q07HA have a significance value > 0.05, which means that they are not significant in predicting a student's AP score. Students who feel proud have the most significant predictor importance, followed by those who feel that their life has a clear meaning or purpose. Both variables indicate that students who suffer from mental health problems tend to have lower AP scores.

C. Performance

To evaluate the performance of each classification model, the top five variables of each factor were fitted into the classification algorithm, except for Demographic, where only one variable is available. Before fitting them into the model, the variables were classified as independent variables, while the AP was classified as the dependent variable. A total of 26 variables were selected and the RapidMiner auto model was used to evaluate the performance of each classification model. The attribute representation for the model construction is shown in Table VI.

TABLE. VI. ATTRIBUTES USED IN THE CLASSIFICATION MODEL

Attribute	Type of Variable	Attributes
26 attributes	Demographic	ST004D01T
	Learning environment at home	ST011Q01TA, ST011Q02TA, ST011Q03TA, ST011Q04TA, ST011Q06TA
	Family background	MISCED, ST123Q02NA, FISCED, PA009Q09NA, EC155Q01DA
	Reading skills and habits	ST013Q01TA, ST154Q01HA, ST161Q06HA, ST152Q05IA, ST153Q06HA
	Career goal and mindset	EC152Q01HA, EC153Q02HA, EC150Q05WA, EC153Q07HA, EC150Q06WA
	Mental health	ST186Q09HA, ST185Q01HA, ST186Q01HA, ST186Q05HA, ST185Q03HA
46 attributes	Demographic	ST004D01T
	Learning environment at home	ST011Q01TA, ST011Q02TA, ST011Q03TA, ST011Q04TA, ST011Q06TA

	Family background	MISCED, ST123Q02NA, FISCED, PA009Q09NA, EC155Q01DA, PA008Q09NA, PA008Q05TA, EC155Q02DA, ST123Q04NA, PA008Q03TA
	Reading skills and habits	ST013Q01TA, ST154Q01HA, ST161Q06HA, ST152Q05IA, ST153Q06HA, ST153Q09HA, ST160Q05IA, ST161Q03HA, ST011Q08TA, ST160Q04IA
	Career goal and mindset	EC152Q01HA, EC153Q02HA, EC150Q05WA, EC153Q07HA, EC150Q06WA, EC150Q07WA, EC153Q01HA, EC150Q01WA, EC153Q05HA, EC153Q06HA
	Mental health	ST186Q09HA, ST185Q01HA, ST186Q01HA, ST186Q05HA, ST185Q03HA, ST186Q02HA, ST186Q10HA, ST186Q08HA, WB171Q03HA, ST186Q06HA

TABLE. VII. PERFORMANCE OF THE CLASSIFICATION MODEL THROUGH RAPIDMINER

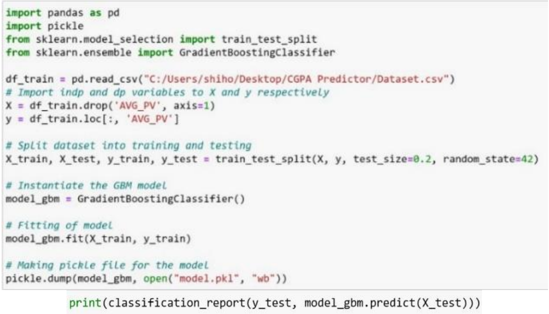
	Accuracy	Precision	Recall
Logistic Regression	72.9%	72.6%	85.9%
Naïve Bayes	72.7%	71.8%	88.1%
Generalized Linear Model	73.0%	71.8%	88.0%
Fast Large Margin	72.5%	71.7%	88.3%
Deep Learning	73.4%	71.4%	91.7%
Decision Tree	65.7%	78.2%	58.0%
Random Forest	71.6%	71.3%	86.6%
Gradient Boosted Trees	73.7%	72.8%	88.3%
Support Vector Machine	72.3%	71.5%	87.4%

Through the analysis of Table VII, it is observed that the prediction of AP using the PISA dataset is quite encouraging for each classification model, with optimistic accuracy, precision, and recall values. This indicates that HEIs may take advantage of the model to predict students' AP. Of all the classification models used, Gradient-Boosted Trees (GBT) outperform other models with the highest accuracy of 73.7% and recall (88.3%-second highest). On the other hand, Decision Tree (DT) performs the worst by obtaining the lowest values for both accuracy and recall. However, DT records the highest precision (78.2%), but RF with the lowest precision (71.3%). This result is consistent with some previous studies (Table I) in which GBT and DT are two models with good performance. However, all models show only slight differences in accuracy and precision.

Before selecting the best classification model to fit the dataset to predict students' AP, each classification model was trained separately on the machine. The code snippet to build the Gradient Boosted Trees is shown in Fig. 3. The independent variables, excluding the PV, were passed into the X variable, while the dependent variable, PV, was passed into the y variable. Then, the X and y variables were split into 80-20 in a random state of 42 to X_train, X_test, y_train, and y_test, where 80% of he data were used for training, and the remaining data were used for testing. Then, the training data (X_train and y_train) were fitted into the classification model, while the test data would be used to evaluate the accuracy, precision, and recall of the model through the classification report. The default GBT parameters were used to avoid overfitting of the data. Since the GBT model will be deployed on a website, the pickle library was used to prevent repetitive training cycles whenever a prediction is to be performed.

The importance of attributes that do not produce higher predictor importance could not be ignored. An additional 20 attributes were fitted to the model taking the top 10 attributes from each category to improve the model accuracy. Only the Gradient-Boosted Trees model will be trained and tested, as it has the highest accuracy out of all the classification models.

When the additional 20 attributes are added, the total attributes to be fitted into the model consist of 46 attributes. The trained model has an accuracy of 73.37%, which shows an increase of 1.64% with the added attributes. Since there are remaining attributes that are not considered in reading skills and habit, career goals and mindset, as well as mental health, attributes with significance < 0.05 will be fitted into the model, except WB171Q04HA, ST186Q07HA and ST186Q03HA. A total of 74 attributes are fitted into the model and have an accuracy of 74.17%. Therefore, there is an increase of 2.44% from the original model with 26 attributes. Table VIII presents the accuracy comparison of the GBT model with different attributes count.



```
import pandas as pd
import pickle
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingClassifier

df_train = pd.read_csv("C:/Users/shiho/Desktop/CGPA Predictor/Dataset.csv")
# Import indp and dp variables to X and y respectively
X = df_train.drop('AVG_PV', axis=1)
y = df_train.loc[:, 'AVG_PV']

# Split dataset into training and testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Instantiate the GBM model
model_gbm = GradientBoostingClassifier()

# Fitting of model
model_gbm.fit(X_train, y_train)

# Making pickle file for the model
pickle.dump(model_gbm, open("model.pkl", "wb"))
```

```
print(classification_report(y_test, model_gbm.predict(X_test)))
```

Fig. 3. Source code for building a gradient booster classifier and evaluation of its performance

TABLE. VIII. ACCURACY OF THE CLASSIFICATION MODEL USING RAPIDMINER

	26 attributes	46 attributes	74 attributes
Gradient Boosted Trees	71.73%	73.37%	74.17%

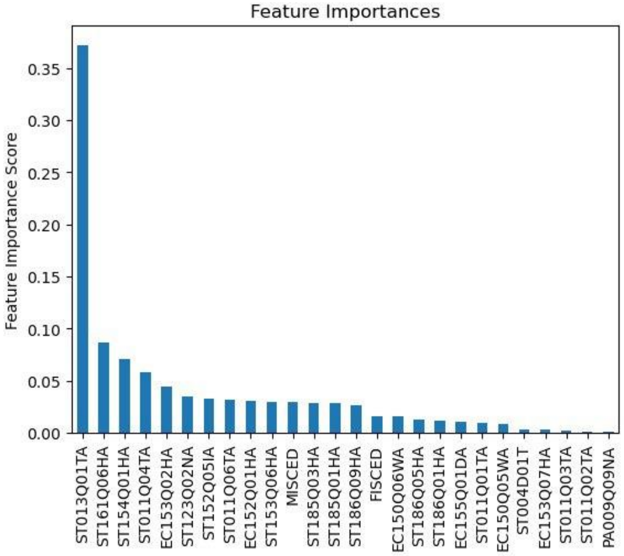


Fig. 4. The feature importance score for each attribute in the original model with 26 attributes

To determine the importance of each factor in predicting academic performance, the feature importance score is evaluated in the GBT model with 26 attributes, which is shown in Fig. 4. Analysis of the characteristic importance score shows that the most significant factor in predicting academic performance is reading skills and habits, which is consistent with the findings of Balan et al. [7]. They claimed that reading cultivates students' critical thinking skills and expands their world views, leading to a positive correlation between reading and academic performance. The top three attributes of the feature importance score are ST013Q01TA, ST161Q06HA, and ST154Q01HA, respectively. This is followed by the learning environment at home - ST011Q04TA, which claimed the second spot among all the factors considered. Career goals and mindset clinched the third spot, EC153Q02HA and subsequently, family history. The findings are supported by a study by Heppt et al. study [62]. They noted that the number of books at home contributes to a child’s understanding of academic language, which aids in academic achievement. However, the information provided by parents on the estimates of the number of books at home is more significant compared to the estimates of children. A detailed overview of each characteristic importance score is illustrated in Fig. 4.

V. Conclusion

This study uses educational data mining techniques to predict the CGPA of TARUMT students based on the PISA dataset from 2018. Through this study, it is established that EDM, which uses machine learning approaches, is constructive in the educational context, enabling institutional leaders to employ predictive modelling techniques to make informed decisions about optimising resource allocation, thus reducing the rate of student attrition rate [63].

The results of this case study revealed that the Gradient Boosted Trees classification technique has the highest performance with 71.7% accuracy, while Decision Tree presents the lowest performance with 59.6%. Furthermore, this case study can increase the completeness in identifying the most significant factors that affect student academic performance. Reading skills and habits, followed by the learning environment at home, which is highly intertwined, are found to have a high correlation with academic performance.

Furthermore, the predictive model resolved external validity concerns using international groups of respondents, which does not restrict itself to a geographical constraint. Therefore, the model could be implemented in different HEIs without the worry of sampling bias.

Using the predictive model, students who are prone to academic failure could be detected, and earlier interventions could be carried out to mitigate their effect. For example, instructors could identify the underlying causes that are likely to cause academic failure toward that student. In doing so, mitigation efforts could be carried out critically, as the root cause of the problem has been identified. Therefore, the tendency to academic failure could be significantly reduced, making academic achievement more easily achievable.

The study had a limitation in that data from the academic results of the students were not used to predict academic performance. Although PVs were supplemented, they may not accurately predict the CGPA of HEIs with more in-depth subject ranges than relying only on high-school-level maths, reading, science, and Global Competency. Although these are fundamental building blocks for furthering students' academic knowledge, a better approach would be to include the CGPA of university students, which covers a wide range of subjects.

The proposed model was trained with students ranging from 15 to 16 years old and tested with students ranging from 20 years old. Therefore, a possibility for future work would be to include university students' data from international groups of respondents. By training the model with data from high school and university students, different inferences could be reached due to the diverse datasets among students at different educational levels.

Another limitation of the study is related to the demographic variables available in the data set. Since the prediction model used data available from the PISA dataset, the school category questionnaire differs from the student questionnaire, where only 10k data is available. This affects the predictor's ability to correctly evaluate the significant importance of demographic factors as only one variable, gender, is used. Therefore, for future work of this study, sufficient data on school category should be included to account for the contribution of demographic variables to predict academic performance.

References

[1] A. Abu Saa, M. Al-Emran, and K. Shaalan, “Factors Affecting Students’ Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques,” Tech Know Learn, vol. 24, pp. 567-598, 2019, doi: 10.1007/s10758-019-09408-7.

[2] S. Al-Sudani and R. Palaniappan, “Predicting students’ final degree classification using an extended profile,” Education and Information Technologies, vol. 24, no. 4, pp. 2357-2369, 2019.

[3] A. Aleryani, W. Wang, and B. Iglesia, “Dealing with Missing Data and Uncertainty in the Context of Data Mining,” in Proceedings of the International Conference on Hybrid Artificial Intelligence Systems, pp. 221-233, 2018.

[4] E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: literature review and best practices,” International Journal of Educational Technology in Higher Education, vol. 17, no. 3, 2020, doi: 10.1186/s41239-020-0177-7.

[5] R. Auerbach et al., “WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders,” Journal of Abnormal Psychology, vol. 127, no. 7, pp. 623-638, 2018, doi: 10.1037/abn0000362.

[6] S. Agnafors, M. Barmark, and G. Sydsjö, “Mental health and academic performance: a study on selection and causation effects from childhood to early adulthood,” Social Psychiatry and Psychiatric Epidemiology, vol. 56, pp. 56, 857–866, 2021, doi: 10.1007/s00127-020-01934-5.

[7] S. Balan, J. E. Katenga, and A. Simon, "Reading Habits and Their Influence on Academic Achievement Among Students at Asia Pacific International University," Abstract Proceedings International Scholars Conference, vol. 7, no. 1, pp. 1490-1516, Oct. 2018, doi: 10.35974/isc.v7i1.928.

[8] B. Bakhshinategh, R. Zaiane, S. Elatia, and D. Ipperciel, “Educational Data Mining Applications and Tasks: A Survey of the Last 10 Years,” Education and Information Technologies, vol. 23, no. 1, pp. 537-553, Jan. 2018, doi: 10.1007/s10639-017-9616-z.

[9] J. Bano, Z. Jabeen, and S. Qutoshi, “Perceptions of Teachers about the Role of Parents in Developing Reading Habits of Children to Improve their Academic Performance in Schools,” Journal of Education and Educational Development, vol. 5, pp. 42-59, Jun. 2018, doi: 10.22555/joeed.v5i1.1445.

[10] J. Barrot, I. Lleanares, and L. Rosario, "Students' online learning challenges during the pandemic and how they cope with them: The case of the Philippines," Education and Information Technology, vol. 26, pp. 7321-7338.

[11] J. Berens, K. Schneider, S. Gortz, S. Oster, and J. Burghoff, "Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods," Journal of Educational Data Mining, vol. 11, no. 3, pp. 1-41, 2019, doi:10.5281/zenodo.3594771.

[12] R. Bruffaerts, P. Mortier, G. Kiekens, R. Auerbach, P. Cujipers, K. Demyttenaere, J. Green, M. Nock, and R. Kessler, "Mental health problems in college freshmen: Prevalence and academic functioning," Journal of Affective Disorders, vol. 225, pp. 97-103, Mar. 2018, doi: 10.1016/j.jad.2017.07.044.

[13] R. Bragança, F. Portela, and M. Santos, "A regression data mining approach in Lean Production," Concurrency and Computation Practice and Experience, vol. 31, no. 1, pp. 1-32, Jan. 2019, doi:10.1002/cpe.4449.

[14] S. Cant, "Hysteresis, social congestion and debt: towards a sociology of mental health disorders in undergraduates," Social Theory and Health, vol. 16, pp. 311-325, 2018, doi:10.1057/s41285-017-0057-y.

[15] K. Chaudhary and N. Gupta, "E-Learning Recommender System for Learners: A Machine Learning based approach," International Journal of Mathematical, Engineering and Management Sciences, vol. 4, pp. 957-967, 2019, doi:10.33889/IJMEMS.2019.4.4-076.

[16] J. Chung and S. Lee, "Dropout early warning systems for high school students using machine learning," Children and Youth Services Review, vol. 96, pp. 346-353, 2019, doi: 10.1016/j.childyouth.2018.11.030.

[17] A. Choi, "Emotional well-being of children and adolescents: Recent trends and relevant factors," OECD Education Working Papers, no. 69, pp. 1-40, 2018, doi:10.1787/41576fb2-en.

[18] G. Crisp, E. Doran, and A. Reyes, "Predicting Graduation Rates at 4-year Broad Access Institutions Using a Bayesian Modeling Approach," Research in Higher Education, vol. 59, pp. 133–155, 2018, doi: 10.1007/s11162-017-9459-x.

[19] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho and G. Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," Journal of Business Research, vol. 94, pp. 335-343, 2018, doi: 10.1016/j.jbusres.2018.02.012

[20] J. García-González and A. Skrita, "Predicting Academic Performance Based on Students' Family Environment: Evidence for Colombia Using Classification Trees," Psychology, Science and Education, vol. 11, no. 3, pp. 299-311, 2019.

[21] W. Grätz and Ø. Wiborg, "Reinforcing at the Top or Compensating at the Bottom? Family Background and Academic Performance in Germany, Norway, and the United States," European Sociological Review, vol. 36, no. 3, pp. 381-394, 2020.

[22] N. Kapasia, P. Paul, A. Roy, J. Saha, A. Zaveri, R. Mallick, B. Barman, P. Das and P. Chouhan, "Impact of lockdown on learning status of undergraduate and postgraduate students during COVID-19 pandemic in West Bengal, India," Children and Youth Services Review, vol. 116, pp. 1-5, 2020.

[23] B. Kumara and B. Kumar, "Impact of Reading habits on the Academic Achievements: A Survey," Library Philosophy and Practice, pp. 1-14, 2019.

[24] T. Le, T. Tran, T. Trinh, C. Nguyen, T. Nguyen, T. Vuong, T. Vu, D. Bui, H. Vuong, P. Hoang, M. Nguyen, M. Ho, and Q. Vuong, "Reading Habits, Socioeconomic Conditions, Occupational Aspiration and Academic Achievement in Vietnamese Junior High School Students," Sustainability, vol. 11, no. 18, pp. 1-29, 2019.

[25] W. Madhoun, "Predictive modelling of student academic performance – the case of higher education in Middle East", Ph.D. thesis, East London Univ., England, 2018.

[26] N. Margaret, "The Relationship between Career Aspiration and Academic Performance of Students in Public Secondary Schools in Nairobi County, Kenya," International Journal of Multidisciplinary Research and Publications, vol. 3, no. 2, pp. 68-73, 2020.

[27] V. Miguéis, A. Freitas, P. Garcia, P and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach", Decision Support System, vol. 115, pp. 36-51, 2018, doi:10.1016/j.dss.2018.09.001.

[28] R. McCarthy, M. McCarthy, W. Ceccucci, and L. Halawi, "Introduction to Predictive Analytics," in Applying Predictive Analytics, Springer International Publishing, 2019, doi:10.1007/978-3-030-14038-0.

[29] F. Okesina, "Influence of Family Background on Academic Performance of Senior Secondary School Students asexpressed by Teachers in Ilorin Metropolis," KIU Journal of Humanities, vol. 3, no. 2, pp. 163-172, 2018.

[30] OECD, "Education at a Glance: OECD Indicators," pp. 1-497, 2019, doi: 10.1787/f8d7880d-en.

[31] A. Oyedeji, A. Salami, O. Folorunsho, and O. Abolade, "Analysis and Prediction of Student Academic Performance Using Machine Learning", JITCE (Journal of Information Technology and Computer Engineering), vol. 4, no. 01, pp.10-15, 2020, doi:10.25077/jitce.4.01.10-15.2020.

[32] B. Prenkaj, P. G. Velardi, D. Distante, and S. Faralli, "A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses," ACM Computing Surveys, vol. 53, no. 3, pp. 1-34, 2021, doi: 10.1145/3388792.

[33] L. Robinson and B. Diale, "Through the eyes of children: Exploring Grade 7 career aspirations," South African Journal of Childhood Education, vol. 7, no. 1, pp. 1-13, 2017.

[34] A. Roy, M. Islam, M. Rahman, M. Saimon, M. Alfaz, and A. Jaber, "A Deep Learning Approach to Predict Academic Result and Recommend Study Plan for Improving Student's Academic Performance," in International Conference on Ubiquitous Computing and Intelligent Information Systems, 2021.

[35] K. Samuel and K. Sylvester, "Read or Perish: Reading Habits among Students and its Effect on Academic Performance: A Case Study of Eastbank Senior High School - Accra," Library Philosophy and Practice, vol. 2018, pp. 1-24, 2018.

[36] N. Sani, A. Nafuri, Z. Othman, M. Nazri, M. Mohamad and N. Khairul, "Drop-Out Prediction in Higher Education Among B40 Students," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 11, pp. 550-559, Nov. 2020, doi: 10.14569/IJACSA.2020.0111169.

[37] G. Sedrakyan, J. Malmberg, K. Verbert, S. Järvelä, and P. Kirschner, "Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation", Computers in Human Behavior, vol. 107, no. C.

[38] Z. Shahbazi, and Y. Byun, "Agent-Based Recommendation in E-Learning Environment Using Knowledge Discovery and Machine Learning Approaches", Mathematics, vol. 10 no. 7, pp. 1-19, 2021.

[39] M. Schippers, and N. Ziegler, "Life Crafting as a Way to Find Purpose and Meaning in Life", Frontiers in Psychology, vol. 10, pp. 1-17, 2019, doi:10.3389/fpsyg.2019.02778.

[40] A. Silva, A. Khatibi, and F. Azam, "Do the Demographic Differences Manifest in Motivation to Learn Science and Impact on Science Performance? Evidence from Sri Lanka," Int. J. Sci. Math. Educ., vol. 16, no. 1, pp. 47-67, Jan. 2018.

[41] F. Ünal, "Data Mining for Student Performance Prediction in Education," in Data Mining - Methods, Applications and Systems, D. Birant, Ed. London: IntechOpen, 2020, pp. 1-21, doi:10.5772/intechopen.91449.

[42] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," Comput. Human Behav., vol. 98, pp. 166-173, May 2019, doi:10.1016/j.chb.2019.04.015.

[43] M. N. Yakubu, and A. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," Kybernetes, vol. 51, no. 2, pp. 916-934, Feb. 2021.

[44] Z. Zainol, and S. Salleh, "Factors Influencing Students Academic Withdrawal during COVID-19 Pandemic," Global Bus. Manag. Res., vol. 13, no. 4, pp. 1-10, Dec. 2021.

[45] Y. Zhao, Q. Xu, M. Chen, and G. Weiss, "Predicting Student Performance in a Master of Data Science Program using Admissions Data," presented at the 13th Int. Conf. on Educational Data Mining (EDM), 2020.

[46] Z. Li and Z. Qiu, "How does family background affect children's educational achievement? Evidence from Contemporary China," J. Chin. Sociol., vol. 5, no. 1, p. 13, 2018, doi:10.1186/s40711-018-0083-8.

[47] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim and P. Garcia, “Using educational data mining techniques to increase the prediction accuracy of student academic performance,” in IEEE Transactions on Education, vol. 62, no. 4, pp. 285-292, Nov. 2019, doi:10.1109/TE.2019.2909471.

[48] A. Shrestha and A. Mahmood, “Review of Deep Learning Algorithms and Architectures,” in IEEE Access, vol. 7, pp. 53040-53065, 2019, doi: 10.1109/ACCESS.2019.2912200.

[49] Z. Pan and M. Cutumisu, “Using machine learning to predict UK and Japanese secondary students' life satisfaction in PISA 2018,” British Journal of Educational Psychology, vol. 94, no. 2, pp. 474-498, 2024, doi:10.1111/bjep.12657 .

[50] S. Acıslı-Celik and C. M. Yesilkanat, “Predicting science achievement scores with machine learning algorithms: a case study of OECD PISA 2015–2018 data,” Neural Computing and Applications, vol. 35, no. 28, pp. 21201-21228, 2023, doi:10.1007/s00521-023-08901-6.

[51] E. G. Bayirli, A. Kaygun and E. Öz, “An Analysis of PISA 2018 Mathematics Assessment for Asia-Pacific Countries Using Educational Data Mining,” Mathematics, vol. 11, no. 6, pp. 1318, 2023, doi:10.3390/math11061318.

[52] J. Chen, Y. Zhang, Y. Wei, and J. Hu, “Discrimination of the contextual features of top performers in scientific literacy using a machine learning approach,” Research in Science Education, vol. 51, no. Suppl 1, pp. 129-158, 2021, doi:10.1007/s11165-019-9835-y.

[53] X. Miao, A. Nadaf and Z. Zhou, “Machine learning evidence from PISA 2018 data to integrate global competence intervention in UAE K–12 public schools,” International Review of Education, vol. 69, no. 5, pp. 675-690, 2023.

[54] A. B. Bernardo, M. O. Cordel, M. O. Calleja, J. M. M. Teves, S. A. Yap, and U. C. Chua, “Profiling low-proficiency science students in the Philippines using machine learning,” Humanities and Social Sciences Communications, vol. 10, no. 1, pp. 1-12, 2023, doi:10.1057/s41599-023-01705-y.

[55] S. Dai, T. Hao, Y. Ardasheva, O. Ramazan, R. W. Danielson, and B. Austin, “PISA reading achievement: Identifying predictors and examining model generalizability for multilingual students,” Reading and Writing, vol. 36, no. 10, pp. 2763-2795, 2023, doi:10.1007/s11145-022-10357-4.

[56] J. Y. Haw and R. B. King, “Understanding Filipino students’ achievement in PISA: The roles of personal characteristics, proximal processes, and social contexts,” Social Psychology of Education, vol. 26, no. 4, pp. 1089-1126, 2023, doi:10.1007/s11218-023-09773-3.

[57] J. Q. Zheng, K. C. Cheung and P. S. Sit, “Identifying key features of resilient students in digital reading: Insights from a machine learning approach,” Education and Information Technologies, vol. 29, no. 2, pp. 2277-2301, 2024, doi:10.1007/s10639-023-11908-0.

[58] Y. Bu and F. Chen, “What key contextual factors contribute to students’ reading literacy among top-performing countries and economies? Statistical and machine learning analyses,” International Journal of Educational Research, vol. 122, pp. 102267, 2023, doi:10.1016/j.ijer.2023.102267.

[59] C. H. Yu, Z. Xiao and J. Hanson, “Machine Learning for Analyzing the Relationship Between Well-Being, Academic Performance with Large-Scale Assessment Data,” Machine Learning in Educational Sciences: Approaches, Applications and Advances, pp. 267-292, 2024, doi:10.1007/978-981-99-9379-6_13.

[60] R. B. King, Y. Wang, L. Fu and S. O. Leung, “Identifying the top predictors of student well-being across cultures using machine learning and conventional statistics,” Scientific Reports, vol. 14, no. 1, 8376, 2024, doi:10.1038/s41598-024-55461-3.

[61] H. M. Low, A. H. L. Lim and F. F. Chua, “Predicting Factors that Affect East Asian Studentsâ€™ Reading Proficiency in PISA,” International Journal on Informatics Visualization, vol. 7, no. 3-2, 2065-2074, 2023, doi:10.30630/joiv.7.3-2.2341.

[62] B. Heppt, M. Olczyk, and A. Volodina, “Number of books at home as an indicator of socioeconomic status: Examining its extensions and their incremental validity for academic achievement,” Soc. Psychol. Educ., vol. 25, pp. 903-928, 2022, doi:10.1007/s11218-022-09704-8.

[63] A. Shanthi, L. T. Heng, E. Sharminnie, P. Purwarno, A. Suhendi, and J. Xavierine, “Do types of gadgets used for online learning have a bearing on student academic performance?,” International Journal of Evaluation and Research in Education (IJERE), vol. 12, no. 4, p. 2222, Dec. 2023, doi: 10.11591/ijere.v12i4.25288.

APPENDIX

APPENDIX A1. VARIABLES USED IN THE PRESENT STUDY

Category	Attribute	Description	Values
Demographic	ST004D01T	Gender	Male; Female
Learning environment at home	ST001Q01TA	A desk to study at	Yes; No
	ST011Q02TA	A personal room	
	ST011Q03TA	A quiet place to study	
	ST011Q04TA ST011Q06TA	A computer Internet connectivity	
	MISCED	Mother’s education level	
	FISCED	Father’s education level	None ISCED 1;ISCED 2;ISCED 3B, C;ISCED 3A, 4;ISCED 5B ISCED 5A, 6
	ST123Q02NA	Parental support	
	PA009Q09NA	Parent’s participation in school	SD-SA
	EC155Q01DA EC155Q02DA	Parent’s participation in schoolwork	Yes;No
	PA008Q09NA	Discuss child’s learning at home with teacher	Never or almost never;A few times a year;About once a month;Several times a month;Several times a week
	PA008Q05TA	Participate in parent council	Yes;No
	ST123Q04NA	Parent’s encouragement to be confident	Yes;No
	PA008Q03TA	Discuss child’s progress with teacher	SD-SA
Reading skills and habits	ST013Q01TA	Number of books at home	Yes;No
	ST154Q01HA	Longest piece of text you read	0–10 books;11–25 books;26–100 books;101–200 books;201–500 books;>500 books
	ST161Q06HA	Difficulty with reading	< 1 page;2 – 10 pages;11 – 50 pages;51 – 100 pages 101 – 500 pages;>500 pages
	ST152Q05IA	Express opinion on a text	SD-SA
	ST153Q06HA	Compare the content of the book or the chapter with your own experience	Never or hardly ever In some lessons;In most lessons;In all lessons
	ST153Q09HA	Select a passage you liked or disliked and explain why	Yes;No
	ST160Q05IA	I read only to get information that I need.	Yes;No

	ST161Q03HA	I read fluently	SD-SA
	ST011Q08TA	Books of poetry	SD-SA
	ST160Q04IA	For me, reading is a waste of time	Yes;No
	ST011Q11TA	Technical reference books at home	SD-SA
	ST011Q07TA	Classic literature at home	Yes;No
	ST153Q01HA	Write a summary of the book or the chapter	Yes;No
	ST160Q02IA	Reading is one of my favourite hobbies.	Yes;No
	ST011Q12TA	Dictionary at home	SD-SA
	ST153Q05HA	Answer questions in class about the book or the chapter	Yes;No
	ST161Q02HA	Ability to answer difficult texts	Yes;No
	ST167Q04IA	Frequency in reading non-fiction books	SD-SA
	ST168Q01HA	Habit in reading books	Never of almost never;A few times a year;About once a month;Several times a month;Several times a week
	ST150Q03IA	Read texts that include tables or graphs	I rarely or never read books; I read books more often in paper format;I read books more often on digital devices. I read books equally often in paper format and on digital devices.
	ST160Q01IA	I read only if I have to.	Many times;Two or three times;Once;Not at all
	ST167Q02IA	Frequency in reading comic books because you want to	SD-SA
	ST153Q03HA	Discuss in small groups with other students who read the same book or chapter	Never of almost never; A few times a year; About once a month; Several times a month, Several times a week
	ST167Q03IA	Frequency in reading fiction because you want to	Yes;No
	ST153Q10HA	Write a text related to what you have read	Never of almost never;A few times a year;About once a month; Several times a month;Several times a week
	ST150Q02IA	Frequency in reading fiction for school	Yes;No
	ST167Q05IA	Frequency in reading newspaper because you want to	Many times;Two or three times;Once;Not at all
	ST153Q04HA	Give personal thoughts about the book	Never of almost never;A few times a year;About once a month;Several times a month;Several times a week
Career goals and mindset	EC152Q01HA	What will you be doing 5 years from now?	Yes;No
	EC153Q02HA	Importance in your close friends' plan for their future	I will be working because the occupation I want does not require a study degree. I will be working because I need to be financially independent I will be studying because I do not know what I would like to do yet. I will be studying because the occupation I want requires a study degree. I will be studying or working for other reasons. I will be doing something else.
	EC150Q05WA	Spoke to career advisor at your school	Not important;Somewhat important;Important;Very important
	EC153Q07HA	Importance in occupational social status.	Yes;No, never
	EC150Q06WA	Discover personal interest and abilities through questionnaire	Not important;Somewhat important;Important;Very important
	EC150Q07WA	Research career information	Yes;No, never
	EC153Q01HA	Importance in parents' expectation on occupation	Yes;No, never
	EC150Q01WA	Did an internship	Not important;Somewhat important;Important;Very important
	EC153Q05HA	Importance in your talents in deciding occupation	Yes;No, never
	EC153Q06HA	Hobbies importance in deciding occupation	Not important;Somewhat important;Important;Very important
	EC150Q09WA	Research internet on future study pathways	Not important;Somewhat important;Important;Very important
	EC153Q04HA	School subjects that you're good at in deciding occupation	Yes;No, never
	EC153Q11HA	Importance of expected salary of occupation	Not important;Somewhat important;Important;Very important
	EC153Q08HA	Importance of financial support for education	Not important;Somewhat important;Important;Very important
Mental Health	ST186Q09HA	How often do you feel proud	Not important;Somewhat important;Important;Very important
	ST185Q01HA	My life has clear meaning or purpose	Never ;Rarely;Sometimes;Always
	ST186Q01HA	How often do you feel joyful	SD-SA
	ST186Q05HA	How often do you feel happy	Never;Rarely;Sometimes;Always
	ST185Q03HA	I have a clear sense of what gives meaning to my life.	Never;Rarely;Sometimes;Always
	ST186Q02HA	How often do you feel afraid	SD-SA
	ST186Q10HA	How often do you feel miserable	Never;Rarely;Sometimes;Always
	ST186Q08HA	How often do you feel sad	Never;Rarely;Sometimes;Always
	WB171Q03HA	Did you feel nervous or tense during a break between classes	Never;Rarely;Sometimes;Always
	ST186Q06HA	How often do you feel scared	Not at all;A little;Quite a bit;Extremely
	ST185Q02HA	I have discovered a satisfactory meaning in life.	Never;Rarely;Sometimes;Always
	ST186Q07HA	How often do you feel lively	SD-SA
	ST186Q03HA	How often do you feel cheerful	Never;Rarely;Sometimes;Always
	WB171Q02HA	Did you feel lonely during a break between classes	Never;Rarely;Sometimes;Always
	WB171Q04HA	Did you feel full of energy during a break between classes	Never;Rarely;Sometimes;Always
	WB171Q01HA	Did you feel happy during a break between classes	Never;Rarely;Sometimes;Always
	ST016Q01NA	Overall, how satisfied are you with your life as a whole these days?	1 - 10